Bayes Homework1

2020311198 Dongook Son

Chapter 1

1.

Conditional probability: suppose that if $\theta = 1$, then y has a normal distribution with mean 1 and standard deviation σ , and if $\theta = 2$, then y has a normal distribution with mean 2 and standard deviation σ . Also, suppose $Pr(\theta = 1) = 0.5$ and $Pr(\theta = 2) = 0.5$.

(a)

For $\sigma = 2$, write the formula for the marginal probability density for y and sketch it.

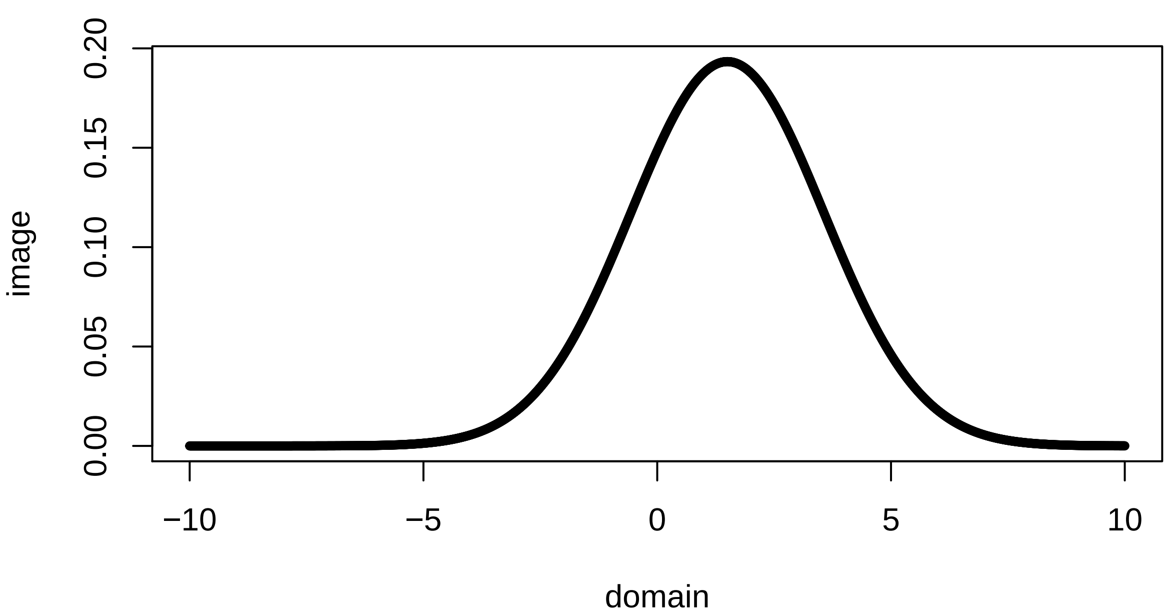
Prior distribution :

$$\pi(\theta) = 0.5 \ where \ \theta \in \{1, 2\}$$

Marginal distribution :

$$\begin{aligned} p(y) &= \pi(\theta = 1)p(y \mid \theta = 1) + \pi(\theta = 2)p(y \mid \theta = 2) \\ &= 0.5\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-1}{2}\right)^2} + 0.5\frac{1}{2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-2}{2}\right)^2} \\ &= \frac{1}{4\sqrt{2\pi}}\left(e^{-\frac{1}{2}\left(\frac{y-1}{2}\right)^2} + e^{-\frac{1}{2}\left(\frac{y-2}{2}\right)^2}\right) \end{aligned}$$

```
domain <- seq(-10,10,0.01)
image <- 0.5*(dnorm(domain,1,2) + dnorm(domain,2,2))
plot(domain, image, cex=0.5)
```



(b)

> What is $Pr(\theta = 1|y = 1)$, again supposing $\sigma = 2$?

By using the Bayes theorem,

$$
\begin{aligned}
Pr(\theta = 1|y = 1) &= \frac{Pr(y = 1|\theta = 1)\pi(\theta = 1)}{\sum_{\theta=1}^{2} Pr(y = 1|\theta)\pi(\theta)} \\
&= \frac{0.5\frac{1}{2\sqrt{2\pi}}}{0.5\frac{1}{2\sqrt{2\pi}}(1 + e^{-\frac{1}{2}(\frac{1-2}{2})^2})} \\
&= \frac{1}{1 + \exp(-1/8)} \\
&= 0.5312
\end{aligned}
$$

```
1/(1+exp(-1/8))
```

```
## [1] 0.5312094
```

(c)

> Describe how the posterior density of θ changes in shape as σ is increased and as it is decreased.

As $\sigma \to \infty$, the posterior density of θ(i.e. $Pr(\theta|y, \sigma)$) converges to 0.5. On the other hand as $\sigma \to 0$, the posterior density converges to 1.
When the variance is vastly large, this implies that the collected data is non-informative hence similar to the prior distribution. But, when the varaince is close to zero it means that the posterior density is a highly decisive function.

2.

> Conditional means and variances: show that (1.8) and (1.9) hold if u is a vector.

Expectation of a vector of random variables stay as a vector, while the variance result in a covariance matrix.

In other words, the expectation can be viewed as a mapping of a vector with n random variables to $\mathbb{R}^n$. Therefore it is obvious that the following holds.

The law of total expectation holds componentwise for $\forall u_i, i \in \{1, 2, ..., n\}$

$$
\begin{aligned}
u = (u_1, u_2, ..., u_n), n \in \mathbb{N} \\
E(u_i) = E(E(u_i|v)) \\
E(u) = E(E(u|v))
\end{aligned}
$$

The variance of a vector, on the other hand, sends n random variables to a $n \times n$ matrix. In other words it is a mapping of $\mathbb{R}^n$ to $\mathbb{R}^n \times \mathbb{R}^n$.

$$
\begin{aligned}
Var(u) &= Cov(u_i, u_j) \ where \ i, j \in \mathbb{N} \\
&= E(u_i u_j) - E(u_i)E(u_j) \\
&= E(u_i u_j) - E(E(u_i|v))E(E(u_j|v)) \\
&= E(u_i u_j) - E(E(u_i|v)E(u_j|v)) \\
&+ E(E(u_i|v)E(u_j|v)) - E(E(u_i|v))E(E(u_j|v)) \\
&= E[E(u_i u_j|v) - E(u_i|v)E(u_j|v)] \\
&+ E(E(u_i|v)E(u_j|v)) - E(E(u_i|v))E(E(u_j|v)) \\
&= E(Cov(u_i, u_j|v)) \\
&+ Cov(E(u_i|v), E(u_j|v))
\end{aligned}
$$

Note that if $i = j$ then $Cov(u_i, u_j) = Var(u_i) = E(Var(u_i|v)) + Var(E(u_i|v))$

3.

Assuming random mating, show that among brown-eyed children of brown-eyed parents, the expected proportion of heterozygotes is 2p/(1 + 2p).

Suppose Judy, a brown-eyed child of brown-eyed parents, marries a heterozygote, and they have n children, all brown-eyed. Find the posterior probability that Judy is a heterozygote and the probability that her first grandchild has blue eyes.

Parents Combination	Parent Pr	Child Pr
2 hetero	$4p^2(1-p)^2$	blue:1/4, hetero:1/2, non-hetero brown:1/4
1 hetero 1 non-hetero brown	$2p(1-p)^3$	hetero:1/2, non-hetero brown:1/2
2 non-hetero brown	$(1-p)^4$	non-hetero brown:1

$$
\begin{aligned}
&P(child : hetero|child : brown, parent : brown) \\
&= \frac{P(child : hetero, child : brown|parent : brown)}{P(child : brown|parent : brown)} \\
&= \frac{P(child : hetero|parent : brown)}{P(child : brown|parent : brown)} \\
&= \frac{0.5 * 4p^2(1-p)^2 + 0.5 * 2p(1-p)^3}{0.75 * 4p^2(1-p)^2 + 2p(1-p)^3 + (1-p)^4} \\
&= \frac{2p}{1+2p}
\end{aligned}
$$

$$
\begin{aligned}
&Pr(Judy : hetero|n - children : brown) \\
&= \frac{Pr(n - children : brown|Judy : hetero)Pr(Judy : hetero)}{Pr(n - children : brown|Judy : hetero)Pr(Judy : hetero) + Pr(n - children : brown|Judy : XX)Pr(Judy : XX)} \\
&= \frac{\frac{2p}{1+2p} * \left(\frac{3}{4}\right)^n}{\frac{2p}{1+2p} * \left(\frac{3}{4}\right)^n + \frac{1}{1+2p} * 1}
\end{aligned}
$$

For Judy's grandchild to have blue eyes, her children must be a heterozygote since all her children have brown eyes.

$$\begin{aligned}
&Pr(Child: hetero|all\ data)\\
&= Pr(Judy: hetero, Child: hetero\ or\ Judy: XX, Child: hetero|all\ data)\\
&= \frac{\frac{2p}{1+2p} * \left(\frac{3}{4}\right)^n}{\frac{2p}{1+2p} * \left(\frac{3}{4}\right)^n + \frac{1}{1+2p}} \left(\frac{2}{3}\right) + \frac{\frac{1}{1+2p}}{\frac{2p}{1+2p} * \left(\frac{3}{4}\right)^n + \frac{1}{1+2p}} \left(\frac{1}{2}\right)
\end{aligned}$$

By the table illustrated above, since Judy's child is a heterozygote, for the grandchild to have blue eyes the spouse must also be a heterozygote.

Spouse	probability	Grandchild:Blue probability
blue	p^2	blue:1/2, hetero:1/2
hetero	$2p(1-p)$	blue:1/4, hetero:1/2, XX:1/4
XX	$(1-p)^2$	hetero:1/2, XX:1/2

$$Pr(Grandchild: blue|\text{all data}) = Pr(child: hetero|\text{all data})*(\frac{1}{4}2p(1-p)+\frac{1}{2}p^2) = Pr(child: hetero|\text{all data})*\frac{p}{2}$$

4.

(a)

Estimate each of these using the relative frequencies of games with a point spread of 8.

$$Pr(\text{favorite wins}|spread = 8) = \frac{\#\text{ of favorite wins with spread 8}}{\#\text{ of outcomes with spread 8}} = \frac{8}{12}$$

$$Pr(\text{favorite wins by at least 8}|spread = 8) = \frac{\#\{13, 15, 16, 20, 21\}}{12} = \frac{5}{12}$$

$$Pr(\text{favorite wins by at least 8}|spread = 8, \text{favorite wins}) = \frac{\#\{13, 15, 16, 20, 21\}}{\#\{1, 6, 7, 13, 15, 16, 20, 21\}} = \frac{5}{8}$$

(b)

Estimate each using the normal approximation for the distribution of (outcome - point spread).

I will use parameters provided from example 1.6.

$$d = \text{outcome - point spread} \sim N(0, 14^2)$$

Note that 'favorite wins' can be interpreted as $outcome + spread > 8$. 0.5 is added due to the continuity correction for discrete data to Normal approximation.

$$Pr(\text{favorite wins | spread=8}) = \Phi(\frac{8 + 0.5}{14})$$

$$Pr(\text{favorite wins by at least 8| spread=8}) = \Phi(\frac{0.5}{14})$$

$$Pr(\text{favorite wins by at least 8| spread=8, favorite wins}) = \frac{Pr(\text{favorite wins by at least 8| spread=8})}{Pr(\text{favorite wins | spread=8})} = \frac{\Phi(\frac{0.5}{14})}{\Phi(\frac{8+0.5}{14})}$$

5.

(a)

> Use any knowledge you have about U.S. politics. Specify clearly what information you are using to construct this conditional probability, even if your answer is just a guess.

The election in the US usually is a competition between the Republican and Democratic candidates. We can set n as the total number of votes in an individual election. Let's think of an extreme case where $n = 2$. Then the probability of election being tied in this case would be 1/2. A tie happening in a voting situation will be as scarce as the number of voters. Therefore we can expand this case to $n : 300,000$ and assume that the election resulting in a tie be $1/300,000$. If we think y as the number of votes for the Republican candidates the relation follows.

$$2 * y = n$$

$$Pr(2 * y - n = 0) = Pr(\text{tie happening})$$

$$Pr(\text{No tie happening}) = 1 - \frac{1}{300000}$$

$$Pr(\text{No tie happening for all elections}) = (1 - \frac{1}{300000})^{435}$$

$$Pr(\text{at least one tie happening for all elections}) = 1 - (1 - \frac{1}{300000})^{435}$$

(b)

> Use the following information: in the period 1900–1992, there were 20,597 congres- sional elections, out of which 6 were decided by fewer than 10 votes and 49 decided by fewer than 100 votes.

We can consider the 49 cases decided by fewer than 100 votes as a tie.

$$Pr(2 * y - n = 0) \sim Pr(|2 * y - n| \leq 100) = \frac{49}{20597}$$

$$Pr(\text{at least one tie happening for all elections}) = 1 - (1 - \frac{49}{20597})^{435}$$

6.

$$\begin{aligned} & Pr(\text{identical twins}|\text{twin brothers}) \\ &= \frac{Pr(\text{identical twins, twin brothers})}{Pr(\text{identical twins, twin brothers}) + Pr(\text{fraternal twins, twin brothers})} \\ &= \frac{\frac{1}{300} * 1/2}{\frac{1}{300} * 1/2 + \frac{1}{125} * 1/4} \end{aligned}$$

$$Pr(\text{fraternal twins, twin brothers}) = Pr(\text{fraternal twins})Pr(\text{twin brothers}|\text{fraternal twins}) = \frac{1}{125} * 1/4$$

$$Pr(\text{identical twins, twin brothers}) = Pr(\text{identical twins})Pr(\text{twin brothers}|\text{identical twins}) = \frac{1}{300} * 1/2$$

8.

(a)

In this case A and B assigning different probabilities on the event that '6' appearing on a fair die is unlikely. Since a fair dice has 6 equal spaces with only the number on the surface varying, two rational persons A and B would both assign the same probability.

$$P_A(E) = P_B(E) = \frac{1}{6}$$

(b)

In an extreme case, when n represents the number of countries participating in the World Cup and E being the event Brazil winning, A would assume $P_A(E) = \frac{1}{n}$ on the grounds that it is a game with fixed outcomes just like rolling a dice. B on the other hand would assign $P_B(E) > \frac{1}{n}$ or $P_B(E) < \frac{1}{n}$ considering the players condition, opponent's strategy and other key variables.

9.

```r
bda_example_1_9_simulation <- function(samples=100){
  ### simulation function for example 1.9
  ir_time <- rexp(samples, 1/10)  # Interarrival time generated from Exp(1/10)

  arrived_time <- c()  # Arrived time
  for(i in 1:samples){arrived_time[i]<-(sum(ir_time[seq(i)]))}

  doctor_time <- runif(samples, 5, 20)  # Doctor time generated from Unif(5,20)

  out_time <- rep(0,samples)  # Leaving time
  out_time[1] <- arrived_time[1] + doctor_time[1]  # First entity
  out_time[2] <- arrived_time[2] + doctor_time[2]  # Second entity
  out_time[3] <- arrived_time[3] + doctor_time[3]  # Third entity
  # Until third entity arrives, there is 0 wait time.

  worker <- rep(0, samples)  # Number of doctors seeing patient
  worker[1] <- 1  # First doctor occupied

  wait_time <- rep(0, samples)  # Wait time

  df = data.frame(ir_time, arrived_time, doctor_time, out_time, worker, wait_time)
  # Dataframe

  # Simulation
  for(i in 2:samples){

  df$worker[i] <- i - sum(df$arrived_time[i] > df$out_time[seq(i-1)])
  # Count working doctors

  if(df$worker[i] <= 3) {
    # No waiting Case
    df$out_time[i] <- df$arrived_time[i] + df$doctor_time[i]
    }

  else {
```

```r
    # Waiting Case
    df$worker[i] <- 3
    df$wait_time[i] <- abs(max(df$arrived_time[i]
                               - df$out_time[seq(i-1)][(df$arrived_time[i]
                                                        - df$out_time[seq(i-1)] < 0)]))
    df$out_time[i] <- df$arrived_time[i] + df$wait_time[i] + df$doctor_time[i]
    }
  }


# Cut off simulation results that exceeds 480 minutes
  res <- df[df$arrived_time <= 420,]
  return(res)
}
```

(a)

> Simulate this process once. How many patients came to the office? How many had to wait for a doctor? What was their average wait? When did the office close?

```r
set.seed(1111)
result <- bda_example_1_9_simulation()
nrow(result)
```

```
## [1] 47
```

47 patients came to the office.

```r
(waited_patients = sum(result$wait_time > 0))
```

```
## [1] 15
```

15 patients had to wait for a doctor.

```r
sum(result$wait_time[result$wait_time > 0]) / waited_patients
```

```
## [1] 2.605127
```

Their average wait was 2.61 minutes.

```r
result$out_time[nrow(result)]
```

```
## [1] 437.8718
```

Officed closed after 437.87 minutes since opening. Approximately 4:18pm.

(b)

> Simulate the process 100 times and estimate the median and 50% interval for each of the summaries in (a).

Simulate 100 times.

```r
# Simulate 100 times and store it in list.
result_list <- list()
for (i in 1:100) {
  result_list[[i]] <- bda_example_1_9_simulation()
}
```

Store statistics for 100 simulations.

```r
num_patients <- c()
waited_patients <- c()
avg_wait_time <- c()
closed_time <- c()

for (i in 1:100) {
  num_patients[i] <- nrow(result_list[[i]])
  waited_patients[i] <- sum(result_list[[i]]$wait_time > 0)
  avg_wait_time[i] = sum(result_list[[i]]$wait_time[result_list[[i]]$wait_time > 0]) /
    waited_patients[i]
  closed_time[i] = result_list[[i]]$out_time[nrow(result_list[[i]])]
}
```

```r
summary(num_patients)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   38.00   43.00   42.48   47.00   59.00
```

```r
summary(waited_patients)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     3.0     5.5     6.1     9.0    21.0
```

```r
summary(avg_wait_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.07777 2.11297 2.86201 3.00751 3.58396 7.03685       3
```

```r
# Note that NA's are simulations without waiting time.
```

```r
summary(closed_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   384.1   417.2   425.1   422.6   430.5   439.0
```